Published online 29 October 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D625–D630***
doi:10.1093/nar/gks992

ECMDB: The *E. coli* Metabolome Database

An Chi Guo[1], Timothy Jewison[1], Michael Wilson[1], Yifeng Liu[1], Craig Knox[1], Yannick Djoumbou[2], Patrick Lo[2], Rupasri Mandal[2], Ram Krishnamurthy[2] and David S. Wishart[1,2,3,*]

[1]Department of Computing Science, [2]Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E8 and [3]National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, Alberta T6G 2M9, Canada

Received August 15, 2012; Revised September 20, 2012; Accepted September 28, 2012

ABSTRACT

The *Escherichia coli* Metabolome Database (ECMDB, http://www.ecmdb.ca) is a comprehensively annotated metabolomic database containing detailed information about the metabolome of *E. coli* (K-12). Modelled closely on the Human and Yeast Metabolome Databases, the ECMDB contains >2600 metabolites with links to ∼1500 different genes and proteins, including enzymes and transporters. The information in the ECMDB has been collected from dozens of textbooks, journal articles and electronic databases. Each metabolite entry in the ECMDB contains an average of 75 separate data fields, including comprehensive compound descriptions, names and synonyms, chemical taxonomy, compound structural and physicochemical data, bacterial growth conditions and substrates, reactions, pathway information, enzyme data, gene/protein sequence data and numerous hyperlinks to images, references and other public databases. The ECMDB also includes an extensive collection of intracellular metabolite concentration data compiled from our own work as well as other published metabolomic studies. This information is further supplemented with thousands of fully assigned reference nuclear magnetic resonance and mass spectrometry spectra obtained from pure *E. coli* metabolites that we (and others) have collected. Extensive searching, relational querying and data browsing tools are also provided that support text, chemical structure, spectral, molecular weight and gene/protein sequence queries. Because of *E. coli's* importance as a model organism for biologists and as a biofactory for industry, we believe this kind of database could have considerable appeal not only to metabolomics researchers but also to molecular biologists, systems biologists and individuals in the biotechnology industry.

INTRODUCTION

Metabolomics involves the systematic study of the small molecule metabolites generated through various cellular or physiological processes (1). As an 'omics' discipline closely connected to proteomics, transcriptomics and genomics, metabolomics is considered one of the cornerstones to systems biology. However, relative to its more mature 'omics' cousins, metabolomics still lags far behind in the development of its software and database infrastructure (2). In an effort to address this informatics deficit, we (and many others) have been steadily developing a set of comprehensive and open access tools to lay a more solid software/database foundation for metabolomics (2–4). In particular, our group has developed several widely used organism-specific or discipline-specific databases, including the Human Metabolome Database (HMDB) (5), the Yeast Metabolome Database (YMDB) (6), DrugBank (7), the Toxin/Toxin-Target database (T3DB) (8) and the Small Molecule Pathway Database (SMPDB) (9). HMDB, T3DB, DrugBank and SMPDB were specifically developed to address the metabolomics, toxicology, pharmacology and systems biology associated with humans (i.e. *Homo sapiens*), whereas YMDB was specifically developed to address the metabolomics and systems biology needs for *Saccharomyces cerevisiae.*

The creation and maintenance of organism-specific metabolomics databases is considered by many to be absolutely essential to the field of metabolomics. This is because each organism has a unique and chemically distinct metabolome. The 'naïve' identification of metabolites, by simple mass matching, for instance, without regard to their origin (organism or man-made), can lead to spurious, humorous or meaningless compound identifications (10). Therefore, as part of our ongoing effort to

*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: david.wishart@ualberta.ca create species-specific metabolomic resources for a number of key model organisms, we have now turned our attention to *Escherichia coli*.

Escherichia coli is a gram-negative, rod-shaped, facultative anaerobic, non-sporulating bacterium that is commonly found in the large intestine or hind gut of mammals. Although most *E. coli* strains are harmless, some serotypes (such as O157 or O104) can cause food poisoning. The most commonly studied strain is K-12, which is the subject of ECMDB. As a model organism for many microbiologists, *E. coli* is probably the most intensively studied prokaryote on the planet. Laboratory strains of *E. coli*, which are very easy to culture and genetically transform, have now become the main workhorse for many molecular biology and protein chemistry laboratories. Indeed, it is fair to say that *E. coli* has played a central role in launching and sustaining today's biotechnology revolution. Being one of the first organisms to be fully sequenced (11) and being particularly amenable to genetic manipulation (12), the sequence, function and interacting partner(s) of a significant proportion of the proteins in *E. coli* are reasonably well known (13). This knowledge is contained in a number of excellent *E. coli*-specific resources, including EcoCyc (14), EchoBase (15), the CyberCell Database (CCDB) (16) and EcoGene (17). In addition to being the 'darling' of molecular biologists, *E. coli* has also long been a favourite for classical biochemists and, more recently, systems biologists. As a result, many essential details of cellular metabolism and numerous metabolic pathways have been worked out through careful biochemical studies and extensive metabolic manipulations of *E. coli*. In fact, *E. coli* (strain K-12) was one of the first organisms to have its metabolism computationally modelled (18) and almost fully reconstructed (19). This work has led to the creation of a number of useful resources containing detailed information on many metabolic pathways and metabolic reactions found in *E. coli*, including KEGG (20), Reactome (21), *i*JO1366 (22) and EcoCyc (14).

While to some, *E. coli* may seem like a 'solved' organism, nothing could be further from the truth. Indeed, as a number of recent *E. coli* metabolomic studies have revealed, all of the major pathway databases, such as KEGG and EcoCyc, are still missing large numbers of *E. coli* metabolites and many important metabolite classes (23–26). Furthermore, some of these databases contain many (100s) exotic compounds that have never been reported in *E. coli* or are only found in genetically manipulated *E. coli* strains (24). For many metabolomics researchers, these errors and omissions are serious. Equally troubling to metabolomics specialists is the fact that none of today's current *E. coli* databases provide detailed metabolite descriptions, intracellular metabolite concentrations, growth or substrate conditions, metabolite physicochemical properties, subcellular locations, reference nuclear magnetic resonance (NMR) or mass spectrometry (MS) spectra or other kinds of data that are routinely needed by those studying *E. coli* metabolism or *E. coli* metabolomics. For metabolomics researchers and industrial biotechnologists using *E. coli* to generate chemical products, these kinds of data need to be readily available, experimentally validated, fully referenced, easily searched and readily interpreted. Furthermore, they need to cover as much of the *E. coli* (strain K-12) metabolome as possible. In an effort to address these shortcomings with existing *E. coli* databases and to create a database more targeted to the needs of *E. coli* (and bacterial) metabolomics, we have developed the *E. coli* Metabolome Database (ECMDB).

DATABASE DESCRIPTION

The ECMDB is a blended bioinformatics–cheminformatics database containing quantitative, analytic or molecular-scale information about *E. coli* (strain K-12) metabolites and their associated properties, pathways, reactions functions, sources, enzymes or transporters. The ECMDB builds on the rich data sets already assembled by such resources as EcoCyc (14), EchoBase (15), CCDB (16), KEGG (20), UniProt (27) and YMDB (6). But it also brings in a large body of independently collected literature data and a significant quantity of experimental data that our laboratory has collected, including assigned NMR and MS spectra of reference compounds and validated metabolite concentrations, to complement this electronic and literature-derived data.

To compile, confirm and validate this comprehensive collection of data, more than a dozen textbooks, several hundred journal articles, nearly 20 different electronic databases and at least 15 in-house or Web-based programs were individually searched, accessed, compared, modified, written or run over the course of the past 12 months. The team of ECMDB contributors and annotators included analytical chemists, microbiologists and bioinformaticians, with dual training in computing science and molecular biology/chemistry.

The ECMDB currently contains >2600 *E. coli* metabolite entries that are linked to >25 000 different synonyms. These metabolites are further connected to some 125 non-redundant pathways and >2800 reactions involving ∼1300 distinct enzymes and ∼300 transporters (from the *E. coli* MG1655 genome). More than 300 compounds are also linked to experimentally acquired 'reference' ^{1}H or ^{1}H/^{13}C spectra, 103 compounds to reference ^{13}C NMR spectra and 320 compounds to reference MS/MS spectra. Intracellular concentration data are also provided for nearly 800 different compounds. Almost all of the data in ECMDB correspond to that for *E. coli* (K-12), although some data, especially concentration data, are taken from *E. coli* isolates such as B or B/K-12 hybrids. Relative to other *E. coli* metabolite/pathway databases, ECMDB is substantially larger and significantly more comprehensive. A detailed comparison of ECMDB to other widely known *E. coli* resources is provided in Table 1.

The ECMDB is modelled closely after the HMDB and YMDB. As a result, it has many of the features found in these databases, including efficient user-friendly tools for viewing, sorting and extracting metabolites, spectra, proteins, pathways or chemical taxonomy information. These are available through the ECMDB navigation bar (located at the top of every ECMDB Web page) that lists seven pull-down menu tabs ('Home', 'Browse', 'Search', 'About', 'Help', 'Download' and 'Contact Us'). To further

Table 1. Comparison of the size and content of different *E. coli*-specific or *E. coli*-containing metabolism/metabolomics databases

Database feature	ECMDB	*i*JO1366	EcoCyc	KEGG
Number of metabolites	2610	1136	2324 (including 1238 not found in wild-type *E. coli*)	1307
Number of data fields	77	8	19	12
Number of NMR spectra	775	0	0	0
Number of MS spectra	4035	0	0	0
Number of external database hyperlinks	19	2	3	5
Concentrations	Yes	No	No	No
Compound descriptions	Yes	No	No	No
Cell locations	Yes	Yes	No	No
Pathways	Yes	No	Yes	Yes
Sequence search	Yes	No	Yes	Yes
Structure search	Yes	No	No	Yes
Molecular weight search	Yes	No	Yes	No
NMR spectral search	Yes	No	No	No
MS spectral search	Yes	No	No	No
Chemical taxonomy	Yes	No	Yes (partial)	No

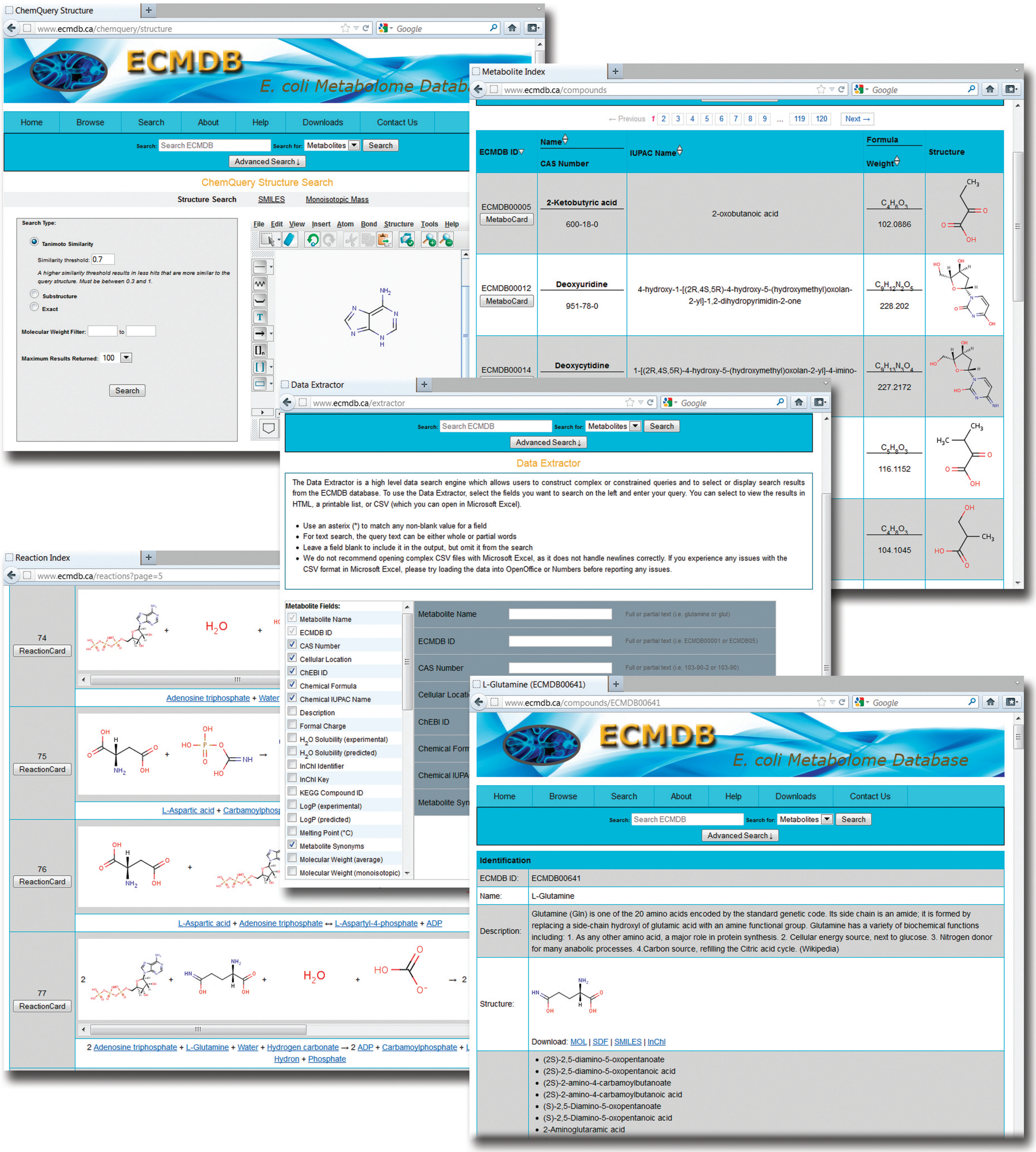
aid in navigation and searching, nearly every viewable page in the ECMDB, including the 'Home' page, supports simple text queries through a text search box located near the top of each ECMDB web page. This text search tool, which can be specified to search through either protein or metabolite data fields, supports text matching, accommodates misspellings and highlights the text where the word is found. A more advanced text search that supports Boolean constructs and permits more precise data field specifications is also available.

In addition to these extensive text search capabilities, the ECMDB also offers general database browsing via the 'Browse' buttons located in the ECMDB menu bar (Figure 1). Five different browsing options are available, including Metabolite Browse (for viewing and sorting metabolites), Protein Browse (for viewing and sorting proteins), Reaction Browse (for viewing chemical reactions), Pathway Browse (for viewing *E. coli*-specific KEGG pathways), Class Browse (for viewing groups of compounds by their chemical taxonomy or class) and Concentration Browse (for viewing and sorting compounds by their measured concentrations). Each of the Browsing views is presented as a set of navigable/sortable synoptic summary tables. These tables are, in turn, linked to more detailed 'MetaboCards' and 'ProteinCards' similar to those found in YMDB and HMDB. Clicking on a MetaboCard or ProteinCard button opens a web page describing the compound or protein of interest in much greater detail. Every MetaboCard entry contains >50 data fields devoted to chemical or physicochemical data, spectral data and synoptic biological data (names, sequences, accession codes). Each ProteinCard entry contains nearly 30 data fields devoted to biochemical, nomenclature, gene ontology and sequence data for metabolically important *E. coli* enzymes and transporters derived from the MG1655 genome sequence. In addition to providing comprehensive numeric, sequence and textual data, each MetaboCard and ProteinCard also contains hyperlinks to many other databases (KEGG, EcoCyc, EchoBase, PubChem, ChEBI, PubMed, PDB, UniProt, GenBank), abstracts, references, digital images and applets for viewing molecular structures and spectra.

Next to the 'Browse' menu, the 'Search' menu offers nine different querying tools, including Chem Query, Text Query, Sequence Search, Data Extractor, MS Search, MS/MS Search, GC/MS search, NMR Search and 2D NMR Search (Figure 1). Chem Query is ECMDB's chemical structure search utility. It can be used to draw a structure (through ChemAxon's freely available chemical sketching applet) or paste a simplified molecular input line entry system (SMILES) string (28) of a query compound into the Chem Query window. Pressing the submit button launches a structure similarity search that looks for common substructures from the query compound that match the ECMDB's database of known *E. coli* compounds. Users can also select the type of search (exact or Tanimoto score) to be performed. High scoring hits are presented in a table with hyperlinks to the corresponding MetaboCards. Chem Query allows users to quickly determine whether their compound of interest is a known *E. coli* metabolite or is chemically related to a known *E. coli* metabolite. In addition to these structure similarity searches, the Chem Query utility also supports compound searches on the basis of molecular weight ranges.

ECMDBs sequence searching utility (Sequence Search) supports single and multiple sequence queries. Specifically, it allows users to use Basic Local Alignment Search Tool to search through ECMDB's collection of 1500 known enzymes, transporters and other target proteins. Using Sequence Search, gene or protein sequences may be searched against ECMDB's sequence database by pasting the FASTA formatted sequence (or sequences) into the Sequence Search query box and clicking the 'submit' button. A significant hit reveals, through the associated MetaboCard hyperlink, the name(s) or chemical structure(s) of metabolites that likely act on that query protein. With Sequence Search metabolite–protein interactions from newly sequenced *E. coli* species or strains may be readily mapped via the *E. coli* data in the ECMDB.

ECMDB's data extraction utility (Data Extractor) uses a simple relational database system that allows users to select one or more data fields and to search for ranges, occurrences or partial occurrences of words, strings or numbers. The data extractor uses clickable Web forms so that users



Search Type:

● Tanimoto Similarity

Similarity threshold: 0.7

A higher similarity threshold results in less hits that are more similar to the query structure. Must be between 0.3 and 1.

○ Substructure

○ Exact

Molecular Weight Filter: to

Maximum Results Returned: 100

ECMDB ID	Name CAS Number	IUPAC Name	Formula Weight	Structure
ECMDB00005 MetaboCard	2-Ketobutyric acid 600-18-0	2-oxobutanoic acid	$C_4H_6O_3$ 102.0886	
ECMDB00012 MetaboCard	Deoxyuridine 951-78-0	4-hydroxy-1-[(2R,4S,5R)-4-hydroxy-5-(hydroxymethyl)oxolan-2-yl]-1,2-dihydropyrimidin-2-one	$C_9H_{12}N_2O_5$ 228.202	
ECMDB00014	Deoxycytidine	1-[(2R,4S,5R)-4-hydroxy-5-(hydroxymethyl)oxolan-2-yl]-4-imino-	$C_9H_{13}N_3O_4$ 227.2172	
			$C_5H_8O_3$ 116.1152	
			$C_4H_8O_3$ 104.1045	

The Data Extractor is a high level data search engine which allows users to construct complex or constrained queries and to select or display search results from the ECMDB database. To use the Data Extractor, select the fields you want to search on the left and enter your query. You can select to view the results in HTML, a printable list, or CSV (which you can open in Microsoft Excel).

- Use an asterix (*) to match any non-blank value for a field
- For text search, the query text can be either whole or partial words
- Leave a field blank to include it in the output, but omit it from the search
- We do not recommend opening complex CSV files with Microsoft Excel, as it does not handle newlines correctly. If you experience any issues with the CSV format in Microsoft Excel, please try loading the data into OpenOffice or Numbers before reporting any issues.

Metabolite Fields:

☑ Metabolite Name
☑ ECMDB ID
☑ CAS Number
☑ Cellular Location
☑ ChEBI ID
☑ Chemical Formula
☑ Chemical IUPAC Name
☐ Description
☐ Formal Charge
☐ H_2O Solubility (experimental)
☐ H_2O Solubility (predicted)
☐ InChI Identifier
☐ InChI Key
☐ KEGG Compound ID
☐ LogP (experimental)
☐ LogP (predicted)
☐ Melting Point (°C)
☑ Metabolite Synonyms
☐ Molecular Weight (average)
☐ Molecular Weight (monoisotopic)

Metabolite Name — Full or partial text (i.e. glutamine or glut)
ECMDB ID — Full or partial text (i.e. ECMDB00001 or ECMDB05)
CAS Number — Full or partial text (i.e. 103-90-2 or 103-90)
Cellular Locati
ChEBI ID
Chemical Form
Chemical IUPA
Metabolite Syn

Adenosine triphosphate + Water

L-Aspartic acid + Carbamoylphosp

L-Aspartic acid + Adenosine triphosphate ↔ L-Aspartyl-4-phosphate + ADP

2 Adenosine triphosphate + L-Glutamine + Water + Hydrogen carbonate → 2 ADP + Carbamoylphosphate + L
Hydron + Phosphate

ECMDB ID:	ECMDB00641
Name:	L-Glutamine
Description:	Glutamine (Gln) is one of the 20 amino acids encoded by the standard genetic code. Its side chain is an amide; it is formed by replacing a side-chain hydroxyl of glutamic acid with an amine functional group. Glutamine has a variety of biochemical functions including: 1. As any other amino acid, a major role in protein synthesis. 2. Cellular energy source, next to glucose. 3. Nitrogen donor for many anabolic processes. 4.Carbon source, refilling the Citric acid cycle. (Wikipedia)
Structure:	Download: MOL | SDF | SMILES | InChI
	• (2S)-2,5-diamino-5-oxopentanoate • (2S)-2,5-diamino-5-oxopentanoic acid • (2S)-2-amino-4-carbamoylbutanoate • (2S)-2-amino-4-carbamoylbutanoic acid • (S)-2,5-Diamino-5-oxopentanoate • (S)-2,5-Diamino-5-oxopentanoic acid • 2-Aminoglutaramic acid

Figure 1. A screenshot montage of ECMDB showing several of the ECMDBs search and data display tools for various metabolites. Not all fields are shown.

may intuitively construct Structured Query Language (SQL)-like queries. Using a few mouse clicks, it is relatively simple to construct complex queries (‘find all metabolites that are substrates of citrate synthase and have melting points >80°C’) or to build a series of highly customized tables. The output from these queries can be provided in Hyper Text Markup Language (HTML) format, with hyperlinks to all associated MetaboCards, or as an easily downloaded comma-separated value (CSV) file.

ECMDB’s NMR and MS search utilities allow users to upload peak lists and to search for matching compounds from the database’s collection of MS and NMR spectra. The ECMDB currently contains 775 experimentally obtained ^1H and ^{13}C NMR spectra (with spectral collection conditions) for ~330 different compounds (most collected in water at pH 7.0, 10 mM for ^1H, 50 mM for ^{13}C) measured in our laboratory or obtained from the BioMagResBank (BMRB) (29). Most of the NMR spectra are fully assigned. It also contains 957 MS/MS (Triple-Quad) spectra for ~320 pure compounds analysed by our laboratory. An additional 3100 MS or MS/MS spectra were obtained from MassBank (30) and Metlin (31). The ECMDB spectral search utilities allow both pure compounds and mixtures of compounds to be identified from their MS or NMR spectra via peak matching algorithms that were developed in-house (32).

Adjacent to the 'Search' menu, the 'About' pull-down menu contains information on the ECMDB database, statistics on *E. coli* (from CCDB), recent news or updates, and links to other databases, data sources and database statistics. The 'Help' pull-down menu provides general documentation on database definitions, data field types and data field sources. It also contains details on how to cite ECMDB as well as a tutorial on how to use ECMDB's advanced text search utilities. Finally, the 'Download' menu contains downloadable data for all ECMDB chemical structures [in Structure Data Format (SDF)], spectral data (in mzML, Chemical Markup Language (CML), NMRSTAR and NMR Markup Language (nmrML) format), all enzyme/protein sequences (in FASTA format) and complete flat file data sets of the current ECMDB release in JavaScript Object Notation (JSON) format.

DATABASE IMPLEMENTATION

ECMDB uses a Ruby on Rails (version: 3.2.3)-based front-end attached to a sophisticated MySQL relational database (version: 5.0.77) at its back-end. All data are entered directly through a custom-built web interface, with each ECMDB MetaboCard having an edit page, which allows database curators to manually make changes to ECMDB entries. The public user interface and the internal database both read from the same database.

All structures and spectra in the ECMDB are stored in centralized structure and spectral hubs. These hubs are Representational State Transfer (RESTful) web resources that automatically store and update spectral properties (assignments, fragmentation assignment data) and chemical properties (molecular weight, solubility and logP). Additionally, these hubs render the spectral and structure images visible on the public ECMDB site. The centralized nature of these spectral and structure hubs helps to maintain consistency for all spectra and structures stored in ECMDB. Whenever a spectrum or a structure is changed or updated, all properties or features are automatically re-calculated and made available on the public site at http://www.ecmdb.ca.

QUALITY ASSURANCE, COMPLETENESS AND CURATION

The same quality assurance, quality control and data compilation procedures implemented during the development of YMDB, HMDB, T3DB and DrugBank were used in the development of ECMDB. In particular, the compounds in ECMDB were identified using a combination of manual literature surveys, text mining of on-line journals or abstracts and data mining of other electronic databases. Literature sources included specialty journals on metabolomics, systems biology, analytical chemistry and textbooks on *E. coli* biochemistry and biology. All metabolites had to have at least two databases to confirm their existence and inclusion (with evidence that the necessary enzymes or pathways are present). For those compounds where there was some ambiguity regarding their source (wild-type K-12 versus genetically modified *E. coli*), we attempted to cross-check our findings through multiple literature sources.

For those *E. coli* metabolites found to match to previously existing entries from either the HMDB or YMDB, only the chemical data fields were imported into the ECMDB (except the compound description, which was manually edited to include or remove organism-specific references). To ensure both completeness and correctness, each metabolite record entered into the ECMDB was reviewed and validated by a member of the curation team after being annotated by another member. Other members of the curation group routinely performed additional spot checks on each entry. Several software packages, including text-mining tools, chemical parameter calculators and protein annotation tools, were developed, modified and used to aid in data entry and data validation. In particular, BioSpider (33) was used extensively to acquire routine, machine retrievable or easily calculated/verifiable chemical data on metabolites. To facilitate and monitor the data entry process, all of ECMDB's data are entered into a centralized password controlled database, allowing all changes and edits to the ECMDB to be monitored, time-stamped and automatically transferred.

CONCLUSION

To summarize, the ECMDB is a richly annotated metabolomics database that brings together quantitative chemical, physical and biological data about 2600 *E. coli* (K-12) metabolites. Relative to other *E. coli* metabolism/pathway databases, ECMDB has between 2–3 times more metabolites and 5–10 times more data. Among the other distinguishing features of ECMDB are (i) the breadth and depth of its annotations (>75 data fields); (ii) the large number of hyperlinks and references to other resources; (iii) the availability of detailed compound descriptions; (iv) the inclusion of thousands of assigned NMR and MS spectral data of reference compounds; (v) the inclusion of intracellular metabolite concentration data; (vi) the quantity of biological and biochemical information included in each compound entry; (vii) the support for queries by text, chemical structure, spectra, molecular weight and gene/protein sequence; and (viii) the support to freely download large components of the database, including sequence, structures and spectra. Because of these unique characteristics, we believe the ECMDB fills an important niche in *E. coli* biology, as it addresses not only the specialized analytical needs of metabolomics researchers but also the interests of molecular biologists, microbiologists, systems biologists and industrial biotechnology.

Although the ECMDB certainly fills an important niche for *E. coli* metabolomics, it is also a work in progress. As with many areas in metabolomics, new compounds are constantly being discovered, new concentrations are being reported, new pathways/reactions are being elucidated and new metabolite functions are being determined. As long as our resources permit, we intend to continue to update and enhance the ECMDB as this new information is published or acquired.

FUNDING

Canadian Institutes of Health Research (CIHR); Genome Alberta, a division of Genome Canada. Funding for open access charge: CIHR.

Conflict of interest statement. None declared.

REFERENCES

1. Vinayavekhin,N., Homan,E.A. and Saghatelian,A. (2010) Exploring disease through metabolomics. *ACS Chem. Biol.*, **15**, 91–103.
2. Wishart,D.S. (2007) Current progress in computational metabolomics. *Brief. Bioinform.*, **8**, 279–293.
3. Wohlgemuth,G., Haldiya,P.K., Willighagen,E., Kind,T. and Fiehn,O. (2010) The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, **26**, 2647–2648.
4. Xia,J., Mandal,R., Sinelnikov,I.V., Broadhurst,D. and Wishart,D.S. (2012) MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.*, **40**, W127–W133.
5. Wishart,D.S., Knox,C., Guo,A.C., Eisner,R., Young,N., Gautam,B., Hau,D.D., Psychogios,N., Dong,E., Bouatra,S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.
6. Jewison,T., Knox,C., Neveu,V., Djoumbou,Y., Guo,A.C., Lee,J., Liu,P., Mandal,R., Krishnamurthy,R., Sinelnikov,I. *et al.* (2012) YMDB: the Yeast Metabolome Database. *Nucleic Acids Res.*, **40**, D815–D820.
7. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
8. Lim,E., Pon,A., Djoumbou,Y., Knox,C., Shrivastava,S., Guo,A.C., Neveu,V. and Wishart,D.S. (2010) T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res.*, **38**, D781–D786.
9. Frolkis,A., Knox,C., Lim,E., Jewison,T., Law,V., Hau,D.D., Liu,P., Gautam,B., Ly,S., Gua,A.C. *et al.* (2010) SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res.*, **38**, D480–D487.
10. Scalbert,A., Brennan,L., Fiehn,O., Hankemeier,T., Kristal,B.S., van Ommen,B., Pujos-Guillot,E., Verheij,E., Wishart,D. and Wopereis,S. (2009) Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, **5**, 435–458.
11. Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
12. Datsenko,K.A. and Wanner,B.L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA.*, **97**, 6640–6645.
13. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al.* (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res.*, **34**, 1–9.
14. Keseler,I.M., Collado-Vides,J., Santos-Zavaleta,A., Peralta-Gil,M., Gama-Castro,S., Muñiz-Rascado,L., Bonavides-Martinez,C., Paley,S., Krummenacker,M., Altman,T. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, **39**, D583–D590.
15. Misra,R.V., Horler,R.S., Reindl,W., Goryanin,I.I. and Thomas,G.H. (2005) EchoBASE: an integrated post-genomic database for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D329–D333.
16. Sundararaj,S., Guo,A., Habibi-Nazhad,B., Rouani,M., Stothard,P., Ellison,M. and Wishart,D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of *Escherichia coli*. *Nucleic Acids Res.*, **32**, D293–D295.
17. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
18. Zeigler,B.P. and Weinberg,R. (1970) System theoretic analysis of models: computer simulation of a living cell. *J. Theor. Biol.*, **29**, 35–56.
19. Feist,A.M., Henry,C.S., Reed,J.L., Krummenacker,M., Joyce,A.R., Karp,P.D., Broadbelt,L.J., Hatzimanikatis,V. and Palsson,B.Ø. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.
20. Kanehisa,M., Goto,S., Saito,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
21. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
22. Orth,J.D. and Palsson,B.O. (2012) Gap-filling analysis of the iJO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions. *BMC Syst. Biol.*, **6**, 30.
23. Ishii,N., Nakahigashi,K., Baba,T., Robert,M., Soga,T., Kanai,A., Hirasawa,T., Naba,M., Hirai,K., Hoque,A. *et al.* (2007) Multiple high-throughput analyses monitor the response of *E. coli.* to perturbations. *Science*, **316**, 593–597.
24. van der Werf,M.J., Overkamp,K.M., Muilwijk,B., Coulier,L. and Hankemeier,T. (2007) Microbial metabolomics: toward a platform with full metabolome coverage. *Anal. Biochem.*, **370**, 17–25.
25. Winder,C.L., Dunn,W.B., Schuler,S., Broadhurst,D., Jarvis,R., Stephens,G.M. and Goodacre,R. (2008) Global metabolic profiling of *Escherichia coli* cultures: an evaluation of methods for quenching and extraction of intracellular metabolites. *Anal. Chem.*, **80**, 2939–2948.
26. Bennett,B.D., Kimball,E.H., Gao,M., Osterhout,R., Van Dien,S.J. and Rabinowitz,J.D. (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.*, **5**, 593–599.
27. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
28. Weininger,D. (1988) SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–38.
29. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
30. Horai,H., Arita,M., Kanaya,S., Nihei,Y., Ikeda,T., Suwa,K., Ojima,Y., Tanaka,K., Tanaka,S., Aoshima,K. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.
31. Smith,C.A., O'Maille,G., Want,E.J., Qin,C., Trauger,S.A., Brandon,T.R., Custodio,D.E., Abagyan,R. and Siuzdak,G. (2005) METLIN: a metabolite mass spectral database. *Ther. Drug Monit.*, **27**, 747–751.
32. Xia,J., Bjorndahl,T.C., Tang,P. and Wishart,D.S. (2008) MetaboMiner—Semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, **9**, 507.
33. Knox,C., Shrivastava,S., Stothard,P., Eisner,R. and Wishart,D.S. (2007) BioSpider: a web server for automating metabolome annotations. *Pac. Symp. Biocomput.*, 145–156.